

Penerapan Algoritma C4.5 dalam Prediksi Potensi Tingkat Kasus Pneumonia di Kabupaten Karawang

Ade Purwanto^{a1}, Aji Primajaya, S.Si., M.Kom^{b2}, Apriade Voutama, M.Kom^{a,b3}

^a *Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Puserjaya, Kec. Telukjambe Timur, Karawang*

¹ade.16010@student.unsika.ac.id

³ apriade.voutama@staff.unsika.ac.id

^b *Universitas Singaperbangsa Karawang*

Jl. HS. Ronggo Waluyo, Puserjaya, Kec. Telukjambe Timur, Karawang

² aji.primajaya@staff.unsika.ac.id

Abstrak

Dilansir dari www.unicef.org disebutkan bahwa hasil analisis baru menunjukkan bahwa, tahun lalu, pneumonia merenggut nyawa lebih dari 800.000 anak balita di seluruh dunia, atau 39 anak per detik. Sebagian besar kematian terjadi pada anak berusia di bawah dua tahun dan nyaris 153.000 kematian terjadi pada bulan pertama kehidupan. Kabupaten Karawang merupakan salah satu Kabupaten yang ada di Jawa Barat. Menurut data rekapitulasi laporan program pengendalian ISPA dinas kesehatan Kabupaten Karawang tahun 2019 terdapat 56,67% anak dibawah 5 tahun yang terdiagnosis penyakit pneumonia dan tingkatan kasus pneumonia terus bertambah setiap tahunnya. *Data mining* merupakan suatu teknik penggalian atau penemuan baru dari data yang besar kemudian mengekstraksi data tersebut menjadi informasi-informasi yang nantinya dapat digunakan. Eksperimen dilakukan dengan menggunakan algoritma C4.5 dan cross validation menggunakan *tools* Weka. Hasil evaluasi yang dilakukan dapat diambil kesimpulan yaitu algoritma C4.5 dengan menggunakan *cross validation* nilai K berbeda-beda memberikan hasil evaluasi yang berbeda-beda. Nilai K=2, K=3, K=4, K=9, K=10 masuk kedalam kategori Good Classification karena mempunyai rentang nilai ROC area yaitu 0,80-0,90 sedangkan untuk nilai K=5, K=6, K=7, K=8 masuk ke dalam kategori *Excelenet Cassification*.

Kata kunci: *Data mining*, Prediksi, Algoritma C4.5, CRISP-DM, Weka, Pneumonia.

Application of C4.5 Algorithm in Prediction of Potential Level of Pneumonia Discovery in Karawang District

Abstract

Reporting from www.unicef.org stated that the results of the new analysis show that, last year, pneumonia claimed the lives of more than 800,000 children under five worldwide, or 39 children per second. Most of the deaths occurred in children under two years of age and nearly 153,000 deaths occurred in the first month of life. Karawang Regency is one of the districts in West Java. According to the data recap of the ISPA control program report, the health department of the district Karawang in 2019 there are 56.67% of children under 5 years of age diagnosed with pneumonia and the rate of pneumonia cases continues to increase every year. Data mining is a technique of extracting or new discoveries from large data and then extracting that data into information that can later be used. Experiments were carried out using the C4.5 algorithm and cross validation using Weka *tools*. The results of the Evaluation can be concluded that the C4.5 algorithm by using the cross-validation of different K values gives different Evaluation results. The value of K = 2, K = 3, K = 4, K = 9, K = 10 is included in the Good Classification category because it has a range of ROC area values, namely 0.80-0.90 while for K = 5, K = 6, K = 7, K = 8 are included in the Excelenet Cassification category.

Keywords: Data mining, prediction, Algorithm C4.5, CRISP-DM, WEKA, Pneumonia

I. PENDAHULUAN

World Health Organization (WHO) menyebut Pneumonia sebagai pembunuh anak-anak terkemuka di seluruh dunia, karena Pneumonia adalah pembunuh utama balita di dunia, lebih banyak dibandingkan dengan

gabungan penyakit AIDS, malaria dan campak. Persentasenya yaitu 19% dari semua penyebab kematian balita, kemudian disusul diare 17% [1]. Data Riset Kesehatan Dasar (Riskesdas) RI tahun 2018 menunjukkan adanya peningkatan prevalensi atau jumlah penderita

pneumonia di bandingkan hasil tahun 2013. Berdasarkan diagnosis tenaga kesehatan jumlah orang yang terkena penyakit ini pada 2013 sekitar 1,6% dan mengalami kenaikan 4% pada tahun 2018. Sedangkan untuk provinsi Jawa Barat pada tahun 2018 mengalami kenaikan sebesar 6% di bandingkan tahun 2013 [2]. Menurut data rekapitulasi laporan program pengendalian ISPA Dinas Kesehatan Kabupaten Karawang tahun 2019 terdapat 56,67% anak dibawah 5 tahun yang terdiagnosis penyakit pneumonia.

Dalam hal ini potensi penyakit pneumonia perlu adanya monitoring secara cepat dan efektif. Dinas kesehatan Karawang selaku badan yang bertanggung jawab dalam mengawasi dan membuat solusi bagi daerah yang terkena penyakit pneumonia. *Data mining* dengan model prediksi mampu membantu dalam memprediksi potensi tingkat kasus pneumonia di Kabupaten Karawang, metode prediksi yang di gunakan dalam penelitian ini yaitu *Decision tree* C.4.5.

Penelitian sebelumnya tentang Komparasi Metode Klasifikasi *Data mining* Algoritma C4.5 dan *Naive bayes* Untuk Prediksi Penyakit Hepatitis. Pada penelitian ini menghasilkan klasifikasi *data mining* Algoritma C4.5 menghasilkan akurasi 77,29% dan nilai AUC 0,846 yang termasuk dalam Good Classification. *Naive bayes* menghasilkan akurasi 83,71% dan nilai AUC 0,812. Dengan demikian dapat disimpulkan bahwa kedua metode ini akurat dalam melakukan prediksi untuk penyakit hepatitis. Algoritma C4.5 lebih unggul dari *Naive bayes* karena memiliki nilai AUC 0,846 dengan kategori *Good Classification* [3].

Penelitian sebelumnya juga dilakukan oleh tentang Perbandingan Algoritma *Naive bayes* dan C4.5 Untuk Klasifikasi Penyakit Anak, Penelitian ini menggunakan *dataset* berupa data rekam medis pasien anak dari Rumah Sakit Islam Yarsi Pontianak sebanyak 240 *record* (kasus) yang terjadi selama tahun 2014 hingga 2016. *Dataset* tersebut terdiri dari 52 atribut berupa gejala penyakit dan menggunakan *cross validation*. Dari hasil perbandingan menunjukkan bahwa algoritma terbaik dengan akurasi sebesar 90.00% yaitu *Decision tree* C4.5. Sementara algoritma *Naive bayes* memperoleh tingkat akurasi sebesar 89.58% [4].

Penelitian lainnya tentang Model Algoritma C4.5 Untuk Prediksi Penyakit Jantung menghasilkan Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy*, dan AUC dari setiap algoritma sehingga didapat pengujian dengan menggunakan C4.5 didapat nilai *accuracy* adalah 86,59 % dengan nilai AUC adalah 0.957, dan masuk kategori kelompok klasifikasi yang sangat baik, karena nilai AUC antara 0.90 sampai 1.00 [5].

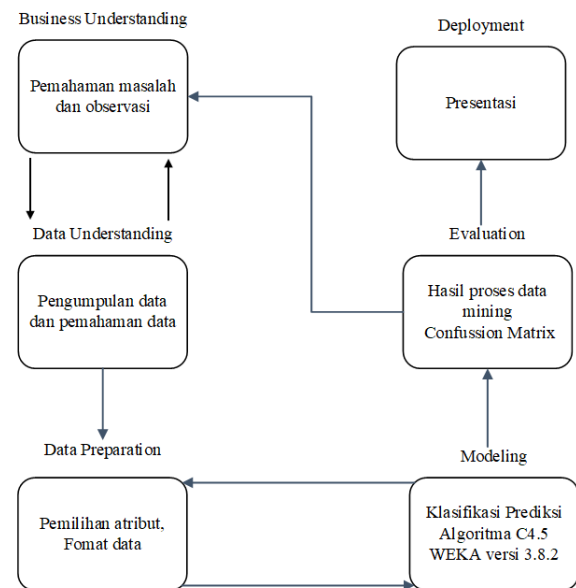
Kelebihan Algoritma *Decision tree* dapat menghasilkan pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik [6].

II. METODOLOGI

A. Metodologi Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini yaitu CRISP-DM (*Cross-Industry Standard Process for*

Data mining). Langkah-langkah yang digunakan yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Untuk menjelaskan rincian langkah-langkah pengolahan CRISP-DM bisa dilihat pada Gambar



Gambar 1. Proses Data mining Metodologi CRISP-DM

1) *Business Understanding*: Pada tahap ini melakukan observasi dan wawancara kepada dinas kesehatan Karawang untuk mengetahui masalah yang terjadi terhadap data penyakit ISPA di Karawang dan melakukan solusi yang tepat untuk permasalahan yang ada.

2) *Data Understanding*: Pada tahap ini dilakukan pengumpulan data yang dibutuhkan yaitu Rekap Laporan Pengendalian ISPA Kabupaten Karawang. Pada tahap ini dilakukan pemahaman tiap atribut yang terdapat dalam data yang sudah diperoleh.

3) *Data Preparation*: Pada tahap ini dilakukan pemilihan atribut atau parameter yang akan di gunakan dalam tahap Modeling. Kemudian melakukan format data karena data yang didapat berupa file Excel, dirubah ke file berformat CSV agar bisa terbaca oleh software WEKA.

4) *Modeling*: Pada tahap ini menentukan teknik data mining yang digunakan, tools data, dan algoritma data mining. Teknik data mining yang digunakan yaitu klasifikasi dengan algoritma C4.5 dan tools WEKA 3.8.2

5) *Evaluation*: Pada tahap ini dilakukan interpretasi terhadap hasil data mining yang ditujukan dalam proses pemodelan pada fase sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama. Evaluasi dilakukan menggunakan Confusion Matrix yang akan menghasilkan menghasilkan nilai *accuracy*, *precision*, dan *recall*.

6) *Deployment*: Pada tahap ini dilakukan penyusunan laporan atau presentasi dari hasil yang didapat dari proses data mining. Laporan akhir melalui pengetahuan yang

didapat atau pengenalan pola pada data dalam proses data mining dan dipresentasikan dalam bentuk grafik atau deskripsi yang mudah dipahami.

B. Studi Litelatur

1) *Pneumonia*: Pneumonia atau dikenal juga dengan istilah paru-paru basah adalah infeksi yang mengakibatkan peradangan pada kantong-kantong udara di salah satu atau kedua paru-paru. Pada penderita pneumonia, sekumpulan kantong-kantong udara kecil di ujung saluran pernapasan dalam paru-paru (alveoli) akan meradang dan dipenuhi cairan atau nanah. Akibatnya, penderita mengalami sesak napas, batuk berdarah, demam, atau menggigil [7].

2) *Data Mining*: Data mining merupakan serangkaian proses dalam pencarian pola, hubungan, penggalian nilai tambah dari data dan informasi yang berukuran besar berupa pengetahuan dengan tujuan menemukan hubungan dan menyederhanakan data agar diperoleh informasi yang dapat dipahami dan bermanfaat dengan bantuan ilmu statistik dan matematika [8]. Data mining merupakan serangkaian proses dalam pencarian pola, hubungan, penggalian nilai tambah dari data dan informasi yang berukuran besar berupa pengetahuan dengan tujuan menemukan hubungan dan menyederhanakan data agar diperoleh informasi yang dapat dipahami dan bermanfaat dengan bantuan ilmu statistik dan matematika [9].

3) *CRISP-DM*: CRISP-DM (*Cross-Industry Standard Process for Data mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri [10]. CRIPS-DM terdiri dari 6 tahap yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *deployment* [11].

4) *Pohon Keputusan*: Pohon keputusan merupakan metode yang umum digunakan untuk melakukan klasifikasi pada data mining. Seperti yang telah dijelaskan sebelumnya, klasifikasi merupakan Suatu teknik menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan [12].

5) *Algoritma C4.5*: Algoritma C4.5 merupakan kelompok algoritma Decision tree. Algoritma ini mempunyai input berupa training samples dan samples. Training samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data. Algoritma C4.5 menghasilkan beberapa aturan-aturan dan sebuah pohon keputusan dengan tujuan untuk meningkatkan keakuratan dari prediksi yang sedang dilakukan, disamping itu

algoritma C4.5 merupakan algoritma yang mudah dimengerti. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi missing value, bisa mengatasi contin data dan pruning. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

- Pemilihan atribut sebagai akar
- Buat cabang untuk tiap-tiap nilai
- Bagi kasus dalam cabang
- Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama [13].

6) *Confusion Matrix*: Confusion matrix adalah alat ukur berbentuk matriks yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi terhadap kelas dengan algoritma yang dipakai [14].

TABEL I
CONFUSION MATRIX

		Nilai Sebenarnya	
		True	False
Nilai prediksi	True	TP	FP
	False	FN	TN

Keterangan:

True Positif (TP) adalah jumlah data positif yang terklasifikasi dengan benar.

True Negative (TN) adalah jumlah data negatif yang terklasifikasi dengan benar.

False Positif (FP) adalah kelas negatif yang terklasifikasi positif.

False Negative (FN) adalah kelas negatif yang terklasifikasi negatif.

7) *WEKA*: WEKA adalah sebuah paket tools machine learning praktis. WEKA merupakan singkatan dari “*Waikato Environment for Knowledge Analysis*”, yang dibuat di Universitas Waikato, New Zealand untuk penelitian, pendidikan dan berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah data mining di dunia nyata, khususnya klasifikasi yang mendasari pendekatan machine learning. Perangkat lunak ini ditulis dalam hirarki class Java dengan metode berorientasi objek dan dapat berjalan hampir di semua platform [15].

III. HASIL DAN PEMBAHASAN

A. Hasil

Hasil dari penelitian *data mining* yang telah dilakukan adalah bagaimana menganalisis algoritma C 4.5 dalam memprediksi potensi tingkat kasus penemonia di Kabupaten Karawang.

1) *Business Understanding*: Tahapan pemahaman bisnis berfokus pada pemahaman tujuan kebutuhan berdasarkan penilaian bisnis. Selanjutnya pemahaman tersebut diubah menjadi sebuah rencana awal data mining untuk mencapai tujuan. Tujuan bisnis ini berdasarkan pada fungsi analisis algoritma C 4.5. tujuannya yaitu agar data jumlah kasus penumonina dapat dimanfaatkan sebaik mungkin dan memberikan informasi untuk memprediksi kasus penumonina di kabupaten Karawang.

Dataset diperoleh dari Dinas Kesehatan Kabupaten Karawang. data ini akan digunakan untuk memperoleh suatu informasi yang dibutuhkan untuk menyelesaikan penelitian ini. Dari informasi yang didapat diharapkan mampu memberikan informasi untuk pemahaman bisnis. Tujuan *data mining* dalam penelitian ini adalah mengetahui model terbaik untuk memprediksi tingkat kasus pneumonia di karawang

2) *Data Understanding*: Tahapan ini merupakan tahap pemahaman data yang diperoleh dari dinas kesehatan Karawang. pemahaman data ini mengacu pada data-data yang telah diperoleh dari Dinas Kesehatan mengenai rekap laporan pengendalian ISPA dinas kesehatan Karawang dimana didalam data terdapat kasus penumonia.

Pengumpulan data yang akan digunakan dalam penelitian ini yaitu data yang didapat dari Dinas Kesehatan Karawang mengenai data mengenai tingkat kasus ISPA di karawang. Keterangan data yang diperoleh bisa dilihat pada TABEL II.

TABEL II
KETERANGAN DATA YANG DIPEROLEH

Nama Data	Tahun	Atribut	Sumber
Rekap Laporan Program Pengendalian ISPA	2016 2017 2018 2019	1) Puskesmas	Dinas Kesehatan Kabupaten Karawang
		2) Jumlah Penduduk	
		3) Jumlah Penduduk usia balita	
		4) Perkiraan Penumonia Balita	
		5) Jumlah kunjungan balita batuk	
		6) Pnumonia < 5 tahun (Laki-laki dan perempuan)	
		7) Penumonia berat < 5 tahun (Laki-laki dan perempuan)	
		8) Pumlah pneumonia < 5 tahun (Laki-laki dan perempuan)	
		9) Jumlah Balita Batuk yg dihitung napas	
		10) Persentase kasus pneumonia <5 tahun	
		11) Batuk bukan Pnumonia	
		12) Jumlah balita batuk yang dihitung napas.	
		13) Presentase yang melaksanakan pemeriksaan dan Tatalaksana Standar	
		14) Jumlah kematian Balita karena pneumonia.	
		15) ISPA >= 5 tahun bukan pneumonia.	
		16) ISPA >= 5 tahun penumonia	

3) *Data Preparation*: Data Preparation mencakup semua kegiatan untuk membangun dataset yang akan dimasukan ke dalam tools pemodelan dari data mentah / awal. Dataset yang diolah dalam pemodelan ini menggunakan algoritma C4.5.

Data yang dipilih adalah data yang diperoleh dari Dinas Kesehatan Kabupaten Karawang antara tahun 2016-2018 untuk data testing dan training sedangkan 2019 untuk uji prediksi hasil model C4.5. Dari dataset yang ada, variabel yang akan di ambil diantaranya adalah Nama puskesmas, jumlah penduduk, jumlah penduduk usia balita, perkiraan penrumonia balita, jumlah kunjungan balita batuk/kesukaran bernafas, jumlah pneumonia < 5 tahun, jumlah penumonia >5 tahun, jumlah kasus pneumonia ditatalaksana standar dan tingkatan kasus penumonia. Tingkatan kasus pneumonia merupakan atribut baru yang ditambahkan yang menyatakan pelabelan dari kelas target. Label merupakan atribut yang menyatakan tingkatan kasus penumonia Rendah, Sedang, Tinggi. Kasus Rendah merupakan label kelas pada data yang memiliki persentase kasus <50%, kasus Sedang merupakan label kelas pada data yang memiliki persentase kasus 50% - 80%, dan kasus Tinggi merupakan label kelas pada data yang memiliki persentase kasus > 80%. Persentase untuk tiap tingkatan kasus berdasarkan arahan dari Dinas Kesehatan Kabupaten Karawang.

Penentuan kelas target dilakukan secara manual oleh peneliti yang diberikan dasar perhitungan oleh Dinas Kesehatan untuk menghitung persentase tingkat kasus penumonia.

$$\text{Persentase kasus} = \frac{\text{jumlah pnemonia}}{\text{perkiraan pnemonia balita}} \times 100\% \quad (1)$$

TABEL II
CONTOH DATA SEBELUM DIBERI KELAS TARGET

Puskesmas	Jumlah Penduduk	Perkiraan Pnumonia Balita	Jumlah Pnumonia < 5 Tahun
Adiarsa	60761	281	220
Anggadita	27095	125	178
Balongsari	18440	85	104
Batujaya	83016	384	307
Bayur Lor	26946	124	143
Ciampel	38077	176	138
Cibuaya	58012	268	286
Cicinde	28954	134	79
Cikampek	104344	482	402
Cikutra	34612	160	81
Tunggakjati	27265	126	40

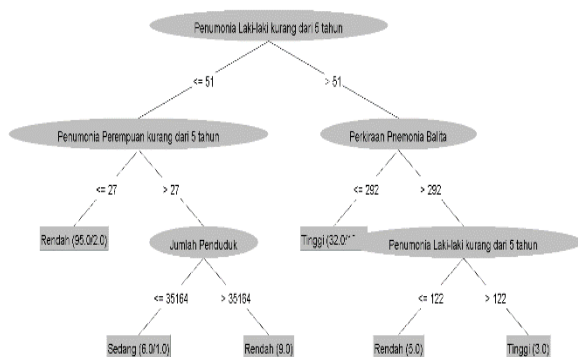
TABEL IV
DATA SETELAH DIBERI KELAS TARGET

Puskesmas	Jumlah Penduduk	Perkiraan Pnemonia Balita	Jumlah Pneumonia < 5 Tahun	Tingkat Kasus
Adiarsa	60761	281	220	Sedang
Anggadita	27095	125	178	Tinggi
Balongsari	18440	85	104	Tinggi
Batujaya	83016	384	307	Tinggi
Bayur Lor	26946	124	143	Tinggi
Ciampel	38077	176	138	Sedang
Cibuaya	58012	268	286	Tinggi
Cicinde	28954	134	79	Sedang
Cikampek	104344	482	402	Tinggi
Cikutra	34612	160	81	Sedang
Tunggakjati	27265	126	40	Rendah

Hasil yang didapatkan dari pemberian kelas target pada data yang digunakan yaitu data tahun 2016-2018 yang berjumlah 150 *record* menghasilkan 107 *record* masuk kedalam kelas target tingkat kasus Rendah, 16 *record* masuk kedalam kelas target tingkat kasus Sedang, dan 27 *record* masuk kedalam kelas target tingkat kasus Tinggi.

4) *Modeling*: Pemodelan adalah fase yang langsung melibatkan data mining. Pemilihan teknik data mining, algoritma dan menentukan parameter dengan nilai yang optimal. Teknik pemodelan data mining yang dipilih adalah klasifikasi prediksi dengan menggunakan algoritma C4.5 decision tree. Klasifikasi algoritma C4.5 digunakan untuk mencapai tujuan awal penelitian yaitu memprediksi tingkat kasus penumonia di Kabupaten Karawang.

Proses pembangunan model algoritma C4.5 menggunakan *tools* weka dan data yang dipakai adalah data rekap laporan ISPA tahun 2016-2018 yang berjumlah 150 *record*. Algoritma C4.5 masuk ke dalam modul *Tree* didalam aplikasi weka dan memiliki kode nama J48. Model yang digunakan yaitu menggunakan algoritma C 4.5 Dalam tahap ini model yang dihasilkan yaitu pohon keputusan yang terdiri dari 6 aturan/rule bisa dilihat pada Gambar 2.



Gambar 2. Pohon Keputusan

Untuk perhitungan secara manual akan dijelaskan sebagai berikut.

Menghitung Entropy (S)

$$\begin{aligned}
 &= \sum_{i=1}^n (-P_i) * \log_2 (P_i) \\
 &= \left(-\left(\frac{107}{150}\right) * \log_2 \left(\frac{107}{150}\right) \right) + \left(-\left(\frac{16}{150}\right) * \log_2 \left(\frac{16}{150}\right) \right) + \left(-\left(\frac{27}{150}\right) * \log_2 \left(\frac{27}{150}\right) \right) \\
 &= 0,347644216 + 0,344407327 + 0,445307614 \\
 &= 1,137359156
 \end{aligned}$$

Menghitung Information Gain

Contoh Menghitung Information Gain (Jumlah Penduduk)

$$\begin{aligned}
 &= Entropy(S) - \sum_{i=1}^k \frac{|S_i|}{|S|} * Entropy(S_i) \\
 &= 1,137359156 - \left(\left(\frac{77}{150}\right) * 1,168134345 \right) - \left(\left(\frac{73}{150}\right) * 1,103016547 \right) \\
 &= 0,000915473
 \end{aligned}$$

Menghitung Split Info

Contoh menghitung Split Info (jumlah penduduk)

$$\begin{aligned}
 &= - \sum_{i=1}^n \frac{S_i}{S} \log_2 \frac{S_i}{S} \\
 &= - \left(\left(\frac{77}{150}\right) * \log_2 \left(\frac{77}{150}\right) \right) + \left(\left(\frac{73}{150}\right) * \log_2 \left(\frac{73}{150}\right) \right) \\
 &= -((-0,49384317) + (-0,505643811)) \\
 &= 0,999486981
 \end{aligned}$$

Menghitung Gain Ratio

Contoh Menghitung Gain Ratio (jumlah penduduk)

$$\begin{aligned}
 &= \frac{\text{information gain (jumlah penduduk)}}{\text{split info (jumlah penduduk)}} \\
 &= \frac{0,000915473}{0,999486981} \\
 &= 0,000915943
 \end{aligned}$$

Dari semua perhitungan nilai entropy dan gain ratio diatas, dapat diketahui bahwa nilai gain ratio terbesar adalah variabel pneumonia laki-laki < 5 tahun dengan nilai 0,564158256 dan yang terkecil adalah jumlah penduduk dengan nilai 0,000915943. Kemudian, ambil nilai gain ratio terbesar yaitu pneumonia laki-laki < 5 tahun dan jadikan sebagai simpul akar. Hilangkan atribut/variabel yang telah dipilih sebelumnya kemudian ulangi perhitungan nilai entriropy dan gain ratio dengan memilih gain ratio terbesar dan dijadikan simpul internal pohon. Ulangi perhitungan tersebut hingga semua variabel pohon memiliki kelas

5) *Evaluation*: Dalam mengevaluasi hasil dilakukan dengan menggunakan cross validation dengan nilai K=2, K=3, K=4, K=5, K=6, K=7, K=8, K=9, K=10. Pengujian evaluasi menggunakan *tools* weka dan *confusion matrix*. Berikut merupakan hasil evaluasi algoritma C4.5 dengan menggunakan cross validation mengenai nilai akurasi, *precision*, *recall*, *f-measure*, serta ROC area ditampilkan pada TABEL IV.

TABEL II
HASIL EVALUASI C 4.5 DENGAN *CROSS VALIDATION*

K=n	Akurasi	Precision	Recall	F-Measure	ROC Area
K=2	84%	0,866	0,840	0,814	0,829
K=3	86%	0,865	0,860	0,862	0,893
K=4	81,3333%	0,806	0,813	0,787	0,860
K=5	84,6667%	0,841	0,847	0,837	0,909
K=6	88%	0,873	0,880	0,875	0,933
K=7	88%	0,879	0,880	0,859	0,926
K=8	86%	0,857	0,860	0,846	0,913
K=9	85%	0,831	0,840	0,822	0,894
K=10	83,3333%,	0,840	0,853	0,842	0,893

Dari TABEL IV dapat diambil kesimpulan dengan menggunakan cross validation nilai K berbeda-beda memberikan hasil evaluasi yang berbeda-beda. Nilai K=2, K=3, K=4, K=9, K=10 masuk kedalam kategori *Good Clasification* karena mempunyai rentang nilai ROC area yaitu 0,80-0,90 sedangkan untuk nilai K=5, K=6, K=7, K=8 masuk ke dalam kategori *exceletnet casification*

6) *Deployment*: Pada tahap ini dilakukan pembuatan laporan hasil atau laporan akhir kegiatan data mining serta dipresentasikan dalam bentuk deskripsi agar lebih mudah dipahami.

B. Pembahasan

Berdasarkan pohon keputusan yang dibangun oleh algoritma C4.5 mengenai prediksi tingkat kasus penumonia di Kabupaten Karawang pada Gambar 2 menghasilkan 6 aturan, yaitu sebagai berikut :

1. JIKA Pneumonia laki-laki kurang dari 5 tahun \leq 51 dan pneumonia perempuan kurang dari 5 tahun \leq 27 MAKA Rendah
2. JIKA Pneumonia laki-laki kurang dari 5 tahun \leq 51 dan pneumonia perempuan kurang dari 5 tahun $>$ 27 dan jumlah penduduk \leq 35164 MAKA Sedang.
3. JIKA Pneumonia laki-laki kurang dari 5 tahun \leq 51 dan pneumonia perempuan kurang dari 5 tahun $>$ 27 dan jumlah penduduk $>$ 35164 MAKA Rendah.
4. JIKA Pneumonia laki-laki kurang dari 5 tahun $>$ 51 dan perkiraan pneumonia balita \leq 292 MAKA Tinggi.
5. JIKA Pneumonia laki-laki kurang dari 5 tahun $>$ 51 dan perkiraan pneumonia balita $>$ 292 dan pneumonia laki-laki \leq 122 MAKA Rendah.
6. JIKA Pneumonia laki-laki kurang dari 5 tahun $>$ 51 dan perkiraan pneumonia balita $>$ 292 dan pneumonia laki-laki $>$ 122 MAKA Rendah.

Uji coba dilakukan menggunakan *tools* Weka dengan *Test-Option Supplied Test Set* karena pengetesan dilakukan dengan menggunakan data lain. Dari hasil percobaan tersebut diketahui jika pemodelan yang dihasilkan algoritma C4.5 dapat memprediksi 50 data baru tersebut

dengan memberikan keterangan berupa tingkat kasus penumonia.

Gambar 3 merupakan hasil dari prediksi.

inst#	actual	predicted	error	prediction
1	1:?	1:Rendah	1	
2	1:?	2:Sedang	0.833	
3	1:?	2:Sedang	0.833	
4	1:?	1:Rendah	0.979	
5	1:?	1:Rendah	0.979	
6	1:?	1:Rendah	0.979	
7	1:?	1:Rendah	1	
8	1:?	1:Rendah	0.979	
9	1:?	1:Rendah	1	
10	1:?	3:Tinggi	0.688	
11	1:?	1:Rendah	0.979	
12	1:?	1:Rendah	0.979	
13	1:?	1:Rendah	0.979	
14	1:?	1:Rendah	0.979	
15	1:?	1:Rendah	0.979	
16	1:?	1:Rendah	0.979	
17	1:?	1:Rendah	0.979	
18	1:?	1:Rendah	0.979	
19	1:?	3:Tinggi	0.688	
20	1:?	3:Tinggi	1	
21	1:?	3:Tinggi	0.688	

Gambar 3. Hasil Prediksi Menggunakan Weka

Hasil prediksi yang didapatkan dari 50 data baru tersebut terdiri dari 39 data untuk tingkat kasus Rendah yang meliputi Adiarsa, Batujaya, Bayur Lor, Ciampel, Cibuaya, Cicinde, Cikampek, Cilamaya, Curug, Gempol, Jatisari, Jayakarta, Jomin, Kalangsari, Karawang, Kutamukti, Kutawaluya, Lemah Duhur, Loji, Majalaya, Medang Asem, Nagasari, Pacing, Pakisjaya, Pangkalan, Pasirukem, Pedes, Plawad, Purwasari, Rngsdengklok, Sukatani, Sungai Buntu, Tanjungpura, Telagasari, Teluk Jambe, Tempuran, Tunggak Jati, Wadas, dan Wanakarta, 3 data untuk tingkat kasus sedang yang meliputi Anggadita, Balongsari, dan Rawamerta, sementara 8 data lainnya masuk kedalam kategori tingkat kasus Tinggi yang meliputi Cikampek Utara, Kertamukti, Klari, Kota Baru, Krawangkulon, Lemah Abang, Tirtajaya, dan Tirtamulya.

IV. KESIMPULAN

Hasil pemodelan algoritma C4.5 dengan menggunakan *tools* Weka menghasilkan pohon keputusan dengan 6 buah rule atau aturan. Untuk mengetahui hasil dari prediksi tingkat kasus pneumonia di Kabupaten Karawang, dilakukan uji coba dengan menggunakan *dataset* dari tahun 2019 yang terdiri dari 50 data. Hasilnya menunjukkan dari 50 data tersebut 39 data masuk kedalam kategori tingkat kasus Rendah, 3 data masuk ke dalam kategori tingkat kasus Sedang dan 8 data masuk kedalam kategori tingkat kasus Tinggi.

Hasil nilai penelitian yang telah dilaksanakan menggunakan Algoritma C4.5 diterapkan dengan metode CRISP-DM menggunakan *tools* Weka 3.8.3. Eksperimen dilakukan dengan menggunakan cross validation. Performa algoritma C4.5 diukur menggunakan nilai akurasi, precision, recall, f-measure, dan ROC area. Hasil evaluasi

yang dilakukan dapat diambil kesimpulan yaitu algoritma C4.5 dengan menggunakan cross validation nilai K berbeda-beda memberikan hasil evaluasi yang berbeda-beda. Nilai K=2, K=3, K=4, K=9, K=10 masuk kedalam kategori Good Classification karena mempunyai rentang nilai ROC area yaitu 0,80-0,90 sedangkan untuk nilai K=5, K=6, K=7, K=8 masuk ke dalam kategori *Excelenet Cassification*.

DAFTAR PUSTAKA

- [1] S. Rahmadhani, M. Jusman Rau, and A. Arifuddin, "Determinan pneumonia pada balita di wilayah kerja puskesmas kamonji kota palu (factors determinants of pneumonia in toddlers at working area of kamonji public health center in palu city)," 2019.
- [2] Kemenkes, "Hasil utama RISKESDAS 2018," Jakarta, 2018.
- [3] W. D. Septiani, "Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.
- [4] S. Bahri, D. M. Midyanti, and R. Hidayati, "Perbandingan Algoritma Naive Bayes dan C4. 5 Untuk Klasifikasi Penyakit Anak," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2018, pp. 24–31.
- [5] A. Rohman and M. Rochcham, "Model algoritma C4.5 untuk prediksi penyakit jantung," *Neo Tek.*, vol. 4, no. 2, 2018.
- [6] A. Rohman and A. Rufiyanto, "Implementasi data mining dengan algoritma decision tree c4. 5 untuk prediksi kelulusan mahasiswa di universitas pandanaran," *Semin. Nas. Teknol. Inf. Apl. Komput.*, pp. 134–139, 2019.
- [7] M. D. C. Pane, "Pneumonia," *Alodoc*, 2020. [Online]. Available: <https://www.alodokter.com/pneumonia>.
- [8] G. Abdurrahman, "Clustering Data Ujian Tengah Semester (UTS) Data Mining," pp. 71–79, 2016.
- [9] G. Abdurrahman, "Clustering Data Ujian Tengah Semester (UTS) Data Mining Menggunakan Algoritma K-Means," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.)*, vol. 1, no. 2, 2016.
- [10] N. R. Syarif and W. Windarto, "Aplikasi Data Mining Dengan Menggunakan Algoritma Fuzzy C-Means dan Metode Recency Frequency Monetary (RFM) Untuk Pengelompokan Pelanggan pada PT Eka Cipta Rasa," *SKANIKA*, vol. 1, no. 3, pp. 1093–1099, 2018.
- [11] N. W. Wardani, *Penerapan Data Mining Dalam Analytic CRM*. Yayasan Kita Menulis, 2020.
- [12] A. Kurniawan, "Implementasi data mining algoritma c4. 5 untuk memprediksi kelulusan uji kompetensi smk teknik komputer dan jaringan (tkj)(study kasus: smk pembangunan daerah lubuk pakam)," *Inf. dan Teknol. Ilm.*, vol. 7, no. 1, 2019.
- [13] K. Umam, D. Puspitasari, and A. Nurhadi, "Penerapan Algoritma C4. 5 Untuk Prediksi Loyalitas Nasabah PT Erdika Elit Jakarta," *J. MEDIA Inform. BUDIDARMA*, vol. 4, no. 1, pp. 65–71, 2020.
- [14] O. Arifin and T. B. Sasongko, "Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier," *Semin. Nas. Teknol. Inf. dan Multimed.* 2018, pp. 67–72, 2018.
- [15] M. F. Aziz, S. Defiyanti, and B. N. Sari, "Perbandingan algoritma cart dan k-nearest neighbor untuk prediksi luas lahan panen tanaman padi di kabupaten karawang," *J. TAM (Technology Accept. Model.)*, vol. 9, no. 2, pp. 74–78, 2019.